Towards Conversational Diagnostic Artificial Intelligence

Team : Asia Ethicists
Jeevan Bhusal
Suhasini Cherukuri
Mithun Thakkar
Hengyuan (Jude) Zhang

Abstract

- Introduce AMIE (Articulate Medical Intelligence Explorer) - LLM based diagnostic dialogue system
- Self-play (simulating its own training data while playing 4 roles – doctor, patient, moderator and critic) and automated feedback (plays the role of critic)
- AMIE performed better as compared to PCPs in terms of accuracy, communication and empathy.
- Limitation - does not replicate real world scenario completely as it is only chat based.

How AMIE Was Evaluated and Key Findings

Evaluation:

- Compared with 20 real primary care physicians.
- Tested 159 cases from Canada, UK, and India.
- Patients were played by trained actors in a text-based consultation format.
- Evaluation done by specialist doctors and patient-actors.

Key Findings:

AMIE performed better than doctors in:

- 30 of 32 categories (specialist rating)
- 25 of 26 categories (patient-actor rating)
- Showed higher diagnostic accuracy, communication skills, and empathy.

Note:

The testing was done using text chat box, it real-world healthcare. in real clinics yet, so further research is needed before real-world use.

Diagnostic Accuracy of AMIE vs PCPs

Higher Diagnostic Accuracy

- AMIE outperforms PCPs across all top-*k* guesses (Top-1 to Top-10)
- Performs better in both disease and non-disease cases
- Slightly lower in obstetrics & gynecology/urology

Robust and Efficient

- Consistent across countries and patient types
- Reasons better with same data
- Collects key info early, with fewer turns

Communication & Consultation Quality

Patient Feedback (N = 46)

- Better than PCPs on 25/26 communication aspects
- High clarity, empathy, professionalism

Specialist Ratings (N = 159)

- Outperformed PCPs on 30/32 points
- Trusted more for diagnosis & management

Self-Evaluation

- AI scores aligned with human experts

Pros and cons in diagnostic and conversational

Advantage

1. Faster response
2. Accurate diagnoses
3. Skillful Communication

Disadvantage

1. Privacy issue
2. Misunderstand human's meaning.
3. Need supervisor
4. Answer are based on prior knowledge

Simulated Dialogue

- Doctor agent does the diagnosis by chatting with patient agent turn by turn. Vignette generator generates patient data. Conversation is performed using simulated dialogue generator.
- Moderator agent decides what is the close of conversation.
- Critic provides feedback to the doctor to improve performance
- Drawback that still remains :–
  - does not model all patient behaviors
  - Moderator assumes that there will always be a diagnosis

Fairness/Bias and Deployment

- To Address Bias:
  - Participatory Evaluation - engaging real people with diverse and underrepresented backgrounds in the AI's evaluation
  - Red-Teaming - involving human or AI Agent evaluators that get deliberately try to fool the system – testing for vulnerabilities
  - Transparent Reporting - Disclosing how the model was trained, what was the data that was used, limitations of the system and what went into the decision making of the system
- Limitations to Deployment
  - There was no process involved wherein if the AI did not perform as expected – a human would be involved.
  - It needs to be updated with most recent clinical knowledge to ensure its usage is reliable for which regulatory readiness and bias auditing are necessary
  - Ensuring having appropriate guardrails to avoid overreliance on the LLM

Conclusion

- Authors conclude by mentioning that AMIE performs well for clinical history taking and empathetic interaction. However, they also say that the current results are only from simulation and a lot more steps need to be taken for real world usage. If this is done – tools like AMIE can be used globally for high-quality healthcare.